Report on the Workshop

"Improving Semantics in Agriculture"

Table of Contents

Prepared by: Tom Baker, Caterina Caracciolo, Yves Jaques
With help from: Erick Antezana, Medha Devare, Lori Finch, Sander Janssen, Philip Roberts, Derek Scuffel

This report: **http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf**
Workshop site: **http://aims.fao.org/agrisemantics-workshop-2015**
Published: Rome: Food and Agriculture Organization of the UN
Date: 29 July 2015

About the workshop

With generous support from the Bill and Melinda Gates Foundation, a workshop was held in July 2015 at the headquarters of the Food and Agriculture Organization of the United Nations in Rome to assess ways of improving the role of "semantics" -- the vocabularies, ontologies, and schemas that give structure to repositories of research results and to agricultural datasets -- in order to strengthen our capacity to meet the challenges of global food security and climate change. For example, semantics can help us identify local varieties of a crop that are more resistant to climate change and to monitor their distribution across multiple language areas. Multilingual semantics can help users find and use data and research produced in different languages. Achieving this requires helping institutions, large and small, organize and publish their data in open and reusable formats.

By bringing together three major agricultural thesauri into a common global repository of conceptual and terminological information, the GACS ("Global Agricultural Concept Scheme") initiative provides a reference point around which we might construct a well-integrated clearinghouse of machine-readable semantic assets in agriculture and nutrition, such as vocabularies, code lists, ontologies, taxonomies, and statistical indicators. This work can benefit researchers, policymakers, extensionists and farmers by offering a backbone of aligned resources that simplify the common tasks of data standardization, comparison, integration, and interpretation. This can lead to accelerated innovation, facilitated academic research, more coherent and interoperable data, and systems and services that help farmers and smallholders become more efficient.

The workshop began with presentations and discussion of ongoing work on modelling and applications in various regions, then split into three breakout groups -- one on issues related to integrating different kinds of semantic assets, one for defining use cases that could benefit from the integration of data through semantically enabled techniques, and one on the state of the art in tools supporting the development of semantically enabled data and the maintenance of related vocabularies (see Appendix A - List of participants, and Appendix B - Agenda). In preparation for the workshop, participants were asked to describe their organizations, vocabularies, datasets, and views on future directions and priorities (see Appendix C). Presentation slides and reports detailing the datasets and vocabularies maintained by participant organizations[1] may be found on the workshop website.[2] This report summarizes the workshop's key findings and conclusions.

1. Integrating semantic assets

Question for discussion: What investments are needed in which areas to bring semantic assets into a common context under Agrisemantics?

Points of agreement

- **Global public goods.** Agrisemantics should provide a mechanism for making semantic assets into global public goods.
- **KOS diversity**. The Agrisemantics platform should encompass many different types of Knowledge Organization Systems (KOS), from formal ontologies and thesauri to code lists ("semantic assets"), in a framework of complimentary specialist vocabularies.
- **Hub-and-spoke model**. N-to-N mappings are inefficient and scale poorly, especially as the number of vocabularies grows. Within Agrisemantics, GACS can play a central role as a hub for more efficient N-to-1 mappings. This is the right time to establish such a hub for agriculture.
- **Distributed maintenance**. Even centralized resources must be maintained in a distributed manner (example: AGROVOC, maintained in over twenty languages).

[1] **http://aims.fao.org/sites/default/files/Agrisemantics_participants.pdf**
[2] **http://aims.fao.org/agrisemantics-workshop-2015**

- **Use of URIs**. The use of URIs to identify terms must be a defining feature of assets within the scope of Agrisemantics. URIs support efficient mapping and the re-use of terms and their labels.
- **Use cases as drivers**. The type of assets to be included under the Agrisemantics umbrella and the priority of mapping activities to be undertaken should be decided based on practical use cases, such as the vocabularies for the management of data on “wheat” studied within the Research Data Alliance (RDA) initiative.

Points of action

1. **Governance**. There is a need for Governing Board to ensure that things get done; a consortium based on an agreement among members; and well-defined process.
2. **Content**. Building on the existing GACS Working Group, there is a need for an Editorial Board.
3. **Infrastructure**. Aside from a website for general information and community interaction, Agrisemantics will need an editorial platform (for maintaining core vocabularies, starting with GACS) and a publication platform (for publishing vocabularies according to Linked Data principles).
4. **Dissemination**. "Telling the story" could involve capacity building, outreach, community workshops, webinars, and tutorials.
5. **Sustainability**. Defining the right organizational form and business model and right-sizing long-term institutional commitments should be a key activity of an Agrisemantics project.
6. **Evaluation**. Semantic Web analytics could provide new ways to measure uptake and impact, and ways should be found to measure benefits to growers and communities.
7. **Use Cases.** See Section 3 below.

Discussion

"Semanticizing" vocabularies. Many if not most of the agricultural vocabularies used within projects and organizations are "pre-semantic" -- that is, they may be locked within a single machine, documented in PDFs or spreadsheets, with no URIs as term identifiers and no defined basis in Semantic Web standards. In addition to providing a publication platform, Agrisemantics should support and build capacity in this area by formulating guidelines and offering tutorials to support the conversion of pre-semantic into semantic vocabularies. Semantic standards should also be used to align the vocabularies of different working domains, such as concept schemes, OWL ontologies, and code lists.

Organizational model. Agrisemantics will be sustainable only if member organizations commit to Agrisemantics as an organizational entity, whether it be a host organization, NGO, or another type of legal entity. Agrisemantics may need a minimal staff, whether seconded from partners or hired with common funds, perhaps on the model of the Gene Ontology consortium.

GACS-centered architecture. Some assets, such as AGROVOC and NALT, can be entirely replaced by GACS. Others, such as the Crop Ontology, or code lists, will only overlap. The idea that GACS can serve as a semantic bridge between diverse types of semantic assets is one that should be scrutinized, tested, and clarified in the context of an architectural study grounded in specific use cases. For example, should GACS relate to application-specific ontologies through direct use of URIs or through mapping? Should cultivars be added directly to GACS Core or linked from a more specialized vocabulary to be made available under Agrisemantics?

Surveying the landscape. The starting scope of Agrisemantics ("version 0.1") should be based on an inventory of available assets. The survey should be used to discover existing linkages between vocabularies already and to prioritize the processes of semanticizing resources and mapping other vocabularies to GACS.

Long-term commitments and sustainability. The institutional commitment of Agrisemantics should be right-sized with regard to long-term services. For example, it might want to guarantee that GACS Core will be actively developed and maintained, while for other assets under the Agrisemantics umbrella it may only want to guarantee the long-term Web persistence of frozen copies (e.g., to ensure that URLs remain resolvable). Accumulating commitments can make Agrisemantics more relevant but also more difficult to sustain. Efficiencies achieved from pooling effort and experience among institutions, as with GACS, can improve sustainability.

Scope of GACS and of Agrisemantics. An Editorial Board will define the scope of GACS Core, with criteria on when to incorporate external terms directly into the core and when to work with other communities on developing more specialized vocabularies that are mapped to the core (the "core-and-outlier" model).

Agrisemantics audience. The Agrisemantics consortium should acknowledge and serve a broadly diverse global community of participants from growers and businesses to researchers and international agencies.

Open access versus licensing. While open access to Agrisemantics assets promotes broad take-up, the project should consider licensing models, such as those deployed in related initiatives such as the European Bioinformatics Institute (EBI) and SNOMED, as a means of achieving sustainability.

Outreach and capacity building. Agrisemantics could have a bigger impact if included in the Knowledge Graph used to generate Google search summaries. Agrisemantics could reach out to the developers of Web and smartphone applications. Webinars and community workshops should describe technologies for exploiting Agrisemantics.

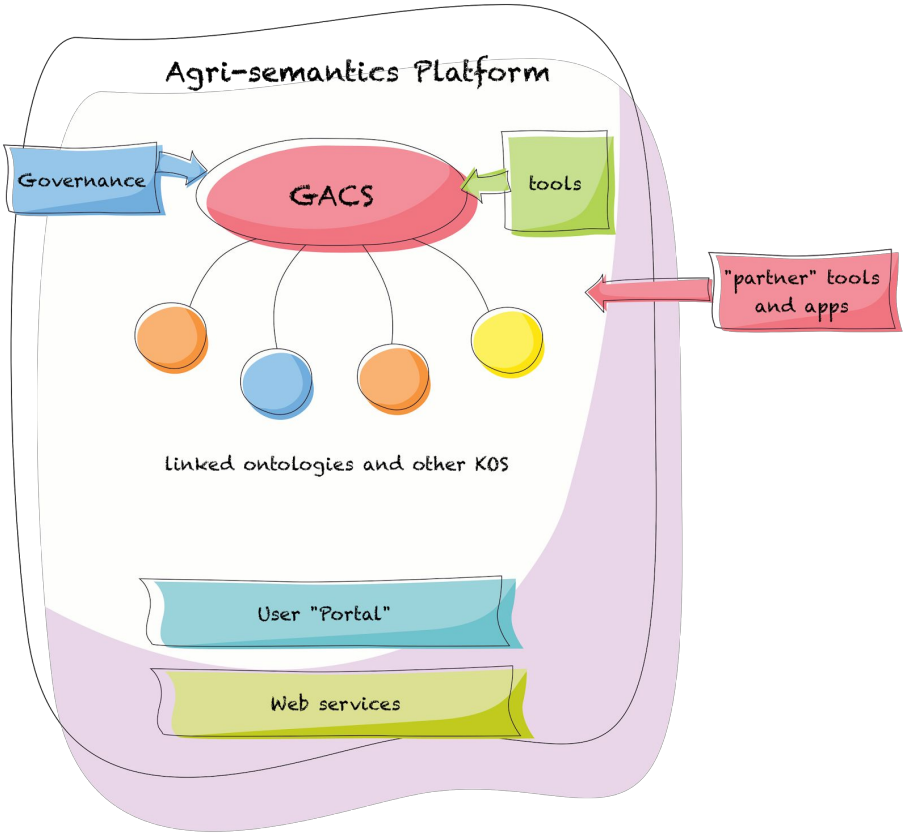

Figure 1. Agrisemantics components as discussed in the breakout sessions

2. Use Cases

Question for discussion: What are some use cases that can help orient a project?

Points of agreement

- **Actors.** A variety of actors are involved in realizing use cases in different roles with different interests; data collectors, data managers, software developers, analysts, researchers, decision makers, extensionists, farmers.
- **Life cycle.** Use cases can potentially involve any part of the data life cycle, from initial data model design, to data gathering, validation, storage, integration, analysis, dissemination, and discovery. Use cases that focus on data usage and visualization often reach the largest number of beneficiaries.
- **Verticality.** Use cases should address a problem in a vertical way, for example in a single country or region, and involve issues and actors along the length of the data life cycle.
- **Quantity.** Three or four use cases can demonstrate the effectiveness of the technology and should focus on impact.
- **Coherence.** Analogous requirements can provide focus and coherence across use cases.
- **Scope.** Use cases should involve more than one problem area -- not just integration or analytic tools but also models, standards, guidelines, good practices, and outreach.
- **Drive and focus:** Use cases must prioritize and drive the work on infrastructure and vocabularies, not the other way around.

Use Case 1: A researcher wants to use multiple datasets for a meta-analysis

Scenario. A researcher specialized in a crop (maize, wheat, beans) wants to integrate fifty datasets for a meta-analysis assessing nutrient use efficiency (example: AgMIP, the Agricultural Model Intercomparison and Improvement Project). A researcher aiming to develop a new variety of wheat that requires fewer inputs will need to survey the literature, then prepare and integrate information from various resources for data mining and further analysis.

Expected benefits are higher wheat crop yields and improved wheat crop resistance. The expected benefits are indirect and dependent on the research. For example, the new variety might take fifteen years to develop, even if the meta-analysis is done in year one.

Building blocks could include the INRA Wheat Initiative's Wheat Information System; AgMIP and the ICASA (International Consortium for Agricultural Systems Applications) standards and wheat modelling group as users; and the Global Agricultural Trial Repository and Database (agtrials.org) of the CGIAR Research Program on Climate Change, Agriculture, and Food Security (CCAFS).

User stories could include:

- As a wheat crop researcher I want to explore impacts to rainfed wheat systems under climate change in Ethiopia so that I can influence policy adaptation discussions in the region.
- As a wheat crop researcher I want to identify wheat genes that control root growth so that I can develop new varieties.

A next step could be to detail ten specific use cases based on different questions that researchers are trying to address and analyse them for common elements.

Potential Product: Wheat Data Monitor

- **Step 1:** Assemble wheat agronomy studies in Ethiopia by searching a variety of literature sources.
- **Step 2**: Extract climate and soil data from those studies.

- **Step 3**: Harmonise the data using the AgMIP format.
- **Step 4**: Conduct a simulation exercise.
- **Step 5**: Analyse outcomes versus previous studies.
- **Step 6**: Visualize results in briefs and communicate with stakeholders.
- **Step 7**: Upload results to Wheat Data Monitor.

Use Case 2: An analyst advises a department head

Scenario. An analyst needs to advise on a policy option or investment in a certain geographic area, taking into account economic, social, and environmental factors.

Expected benefits are better decisions on agricultural issues of benefit to civilians and farmers or, in the case of companies, of benefit to employees and shareholders.

Building blocks could include aid investment data from the International Aid Transparency Initiative (IATI) and CGIAR spending programs; remote sensing data; the ARIES platform for ecosystem assessment based on semantic technologies (ariesonline.org), the Global Yield Gap Atlas, and FAO's Global Agro-Ecological Zones (GAEZ).

User stories could include:

- As an analyst I want to identify segments of farmers who benefitted from development funds for irrigation technology through improved crop productivity so that our donors can improve the way they provide aid.
- As an analyst I want to find out who is doing what where with Wheat Initiative resources so that the Wheat Coordinator can better understand the impact of the initiative.

Potential Product: Agricultural Technology Adoption Tracker (ATAT)

- **Step 1**: Analyst selects relevant indicators (e.g., rainwater harvesting structures and average daily rainfall) and region of interest (e.g., Ethiopia) and requests system to "map it".
- **Step 2**: Analyst adds indicators (e.g., maize productivity) and system presents results visually or in tabular form.
- **Step 3**: Analyst requests system to run correlation analysis on maize productivity and presence or absence of rainwater harvesting structures.
- **Step 4**: Analyst assesses the impact of the irrigation technology on maize productivity.

Key aspects

- **Annotated datasets.** The product would annotate perhaps twenty datasets in order to improve usability and applicability, using tags from perhaps five vocabularies.
- **Datasets as public goods.** The product would make basic datasets on climate, soil, yield gaps, land use, investment aid, biodiversity, and water resources available as public goods in a form that is maintainable and extensible in the future.
- **Interoperability between vocabularies.** The product would require interoperability between vocabularies that address different purposes.
- **Spatial and temporal references.** All data would be identified by at least these two key attributes.
- **Outreach and advocacy.** The product would involve educating data suppliers on the role and use of standards and vocabularies in the context of tools supporting data collection and use.

Use Case 3: An information service provider advises farmers

Scenario. An information service provider needs to transform complex science and technology datasets and information analyses to digestible pieces in order to provide advice to farmers using tools with low-usability barriers. (This scenario received the most votes in the closing plenary discussion.)

Expected benefits are better farm management, with lower losses for farmers.

User Stories

- As an extensionist I want to discover and promote a portfolio of feasible conservation agriculture techniques so that I can help my local farmers improve soil fertility.
- As an extensionist I want to discover information on how to deal with a major pest outbreak in my area so that I can give advice and recommendations to my local farmers.

Building blocks could include the research literature as indexed in AGRIS, Agricola, and CAB Abstracts, the Plantwise program, the Famine Early Warning Systems Network (FEWS-NET), the CIMMYT RustMapper, and possibly the ATAT product mentioned in Use Case 2.

Potential Product: Recommender

Generates recommendations based on research literature, using vocabularies to inter-link data sources, and supports the creation of Early Warning Systems. The steps below would be performed as part of an Agrisemantics project and not by the end user as in the previous use case.

- **Step 1**: Current vocabularies used by building blocks listed above are extended and cross-linked to meet the needs of the recommender.
- **Step 2**: Vocabularies enable consistent recommendations.
- **Step 3**: Learning system is implemented with a reality checker to improve relevance of results.
- **Step 4**: Learning system is localized for the area where it will be tested and system is made available also in offline mode.
- **Step 5**: A feedback mechanism from service providers and farmers back into system is created for continual improvement.

Use Case 4: An IT developer publishes open data

Scenarios. An IT developer is charged with pragmatically implementing institutional policy on open data and needs an infrastructure of simple and open tools tailored for easy use in agriculture, along with capacity building measures.

Expected benefits are increased discoverability, comparability, and usability of organizational data.

User stories

- As a statistical time-series data manager I want to publish our organization's data using open standards so that I can make our data as easy to discover and use as possible.
- As a software developer creating data mashups I want to easily mash up our organization's data so that I can help external users easily navigate across different data sets.

- As a donor I want to be able to easily combine data from recipient organizations so that I can easily compare results.

Building blocks potentially include FAO CountryStat, CGIAR Investment Mapping Tool, the Comprehensive Knowledge Archive Network (CKAN), the Tabular Data Working Group, Statistical Data and Metadata Exchange (SDMX), the Data Documentation Initiative (DDI), Extended Knowledge Organization System (XKOS), Generic Statistical Information Model (GSIM), the CGIAR Data Management Support Pack, and Africa Information Highway (AIH).

Potential product: Tools, Standards and Guidelines

- Focus on one or two institutions in one or two countries.
- Integrate a toolbox for data manipulation.
- Create training webinars, cheat sheets, and FAQs on data management.

General discussion of use cases

- Use cases must consider the business processes behind the data life cycle, from collection to usage.
- Use cases should consider not only tools but also outreach.
- The example use cases described here are merely points of departure. Many more use cases are possible and could potentially be submitted by workshop participants.
    - The group cautioned that potential difficulties must not be ignored.
    - The group started to elaborate how semantic technologies might benefit the use cases but this aspect would require further work.
- As a next step, priorities should be set for selecting the use cases. Do the use cases have commonalities? Which are out of scope or too complex? In a poll of all workshop participants, Use Case 3 was clearly preferred, followed by 2 and 4 closely together, and with the least votes for 1.
- The resources required for outreach and capacity development would need to be considered.
- Need to develop and identify curricula for data scientists and others involved in the data life cycle.
- Need to reach out to existing cross-domain standards bodies such as SDMX, OGC and DDI.

3. Tools

The group discussion emphasized the gap still to be filled in a number of areas. In particular, tools should be developed, as needed, to support the following actions:

- the refinement (as necessary) and reuse of vocabularies and ontologies, to avoid repeatedly re-creating the same resources;
- the use and installation by IT people of different background, not only specifically “semantic.” In this respect, the important of documentation cannot be over emphasized;
- the use of resources that lie along the whole spectrum of semantic vocabularies, not only ontologies and thesauri, but also controlled lists and lightweight ontologies;
- the semantic description not only of ontologies and vocabularies, but of datasets. In particular, tools for the automatic production of metadata are needed;
- to cope with widespread tech-aversion of average end users, which makes them usually unwilling to install new technologies in their institutions, or to spend extensive amount of time to learn new tools;
- to perform “model transformation” and lifting to more expressive languages.

The groups discussed the issue of sustainability of the whole infrastructure that should support the actions mentioned above. Broad agreement was reached that the infrastructure should be run at the community level,

possibly with a distributed architecture of both pre-existing and new services connected into a network. At the level of a single institution, any project that aims at sustainability should take into account the lack of IT hardware resources, as well as the lack of trained IT personnel, especially in the case of small institutions and especially in developing countries.

Tools currently in use could be shared. The Agrisemantics architecture must be developed in stages, progressively involving organizations in a common infrastructure, not just "pushing" to the agrisemantics platform but also "pulling" data. This can only occur fairly with a sharing of all partner knowledge and case studies.

Conclusion

The AgriSemantics workshop convened experts from a wide range of agricultural fields under the common topic of "semantics" in data management. It is uncommon to find managers of soil maps, bibliographic data on research, and plant genome databases in the same room. The presentations introduced the state of the art in each community and identified common issues. In two very full days, the group moved closer to a shared vision of follow-up projects:

- **Breakout Group 1** emphasized the importance of semantic assets -- concept schemes, ontologies, code lists, and the list -- as global public goods. These diverse data vocabularies could coalesce around a hub, the Global Agricultural Concept Scheme (GACS), in a distributed, URI-based model, with a governance mechanism and editorial board.
- **Breakout Group 2** viewed semantic interoperability as a challenge spanning the entire life cycle of data, from data design and collection to archiving. Projects must address the needs not just of end users, but of other actors such as analysts, who need more efficient ways to integrate data for meta-analyses, and data managers, who need to publish data for easy consumption.
- **Breakout Group 3** acknowledged a need for documentation and outreach to data providers on the benefits of shared semantic standards. The processes of maintaining semantic datasets, creating semantic descriptions (metadata), and "semanticizing" existing data vocabularies must be supported by tools that are usable in low-resource conditions and could, potentially, be run and supported at the community level with a distributed architecture of services.

Following up on the workshop, CGIAR and FAO will produce a concept paper for a use-case-driven project. The concept will strive to address the points of agreement that emerged from the breakout groups, focusing on a consortium driven effort to align concept schemes, ontologies and code lists in order to concretely implement the work within one or more existing data-driven information systems, enhancing their efficiency and increasing their value to end users.

Appendix A - List of Participants

Erick Antezana (Bayer CropScience) [1,4]
Elizabeth Arnaud (Bioversity)
Ioannis Athanasiadis (Democritus University of Thrace) [2]
Sophie Aubin (INRA) [3]
Pascal Aventurier (INRA) [1]
Melanie Bacou (IFPRI) [2]
Tom Baker (consultant and Sungkyunkwan University) [1,4]
Hugo Besemer (Wageningen UR) [1]
Christopher Brewster (Aston University of Bologna) [2]
Caterina Caracciolo (FAO) [1,4]
Lucy Chang (FAO)
Medha Devare (CGIAR) [2,4]
Anton Doroshenko (CABI UK) [1]
John Fereira (Cornell University) [3]
Lori Finch (USDA National Agricultural Library) [1,4]
Sander Janssen (Alterra, Wageningen UR) [2,4]
Yves Jaques (FAO) [2,4]
Esther Dzale Yeumo Kabore (INRA) [2]
Richard Kedemi (KARI) [3]
Johannes Keizer (AIMS) [1]
Giovanni L'Abate (CRA) [2]
Simon Liu (USDA) [1]
Devika Madalli (Institute of Statistics India) [1]
Thembani Malapela (FAO) [3]
Daniel Martini (KTBL) [3]
Laura Meggiolaro (IFAD)
Ruthie Musker (University of California at Davis) [1]
Cristina Perez (IFPRI) [1]
Ivo Pierozzi (Embrapa) [2]
Philip Roberts (CABI UK) [1,4]
Kate Schneider (Bill and Melinda Gates Foundation) [2]
Derek Scuffel (Syngenta) [1,4]
Rosemary Shrestha (CIMMYT) [1]
Armando Stellato (University of Rome Tor Vergata) [3]
Giannis Stoitsis (AgroKnow) [3]
Imma Subirats (RDA)
Marc Taconet (FAO Fisheries)
Tom Tomich (University of California at Davis)
Ferdinando Villa (Basque Centre for Climate Change) [1]
Karna Wegner (FAO)
Jeffrey White (USDA Agricultural Research Service) [2]

[1] Participant in Group 1 on "interlinked semantics".
[2] Participant in Group 2 on "use cases".
[3] Participant in Group 3 on "tools".
[4] Co-author of this workshop report

Appendix B - Workshop agenda

Thursday, 2 July 2015

Opening session

- Context and goals for the workshop (Johannes Keizer, FAO and Kate Schneider, Gates)
- Summary report based on pre-workshop questions (Caterina Caracciolo, FAO)

Modelling (Plenary)

- Integrated Modelling on the Semantic Web (Ferdinando Villa, Basque Centre for Climate Change)
- Crop Ontology: Harmonizing Breeders' Data (Elizabeth Arnaud, Bioversity)
- The Global Agricultural Concept Scheme (GACS) project (Tom Baker)
- Sustainable Sourcing of Agricultural Raw Materials (Ruthie Musker, UC Davis)

Applications (Plenary)

- Soil Vocabularies and Applications (Giovanni L'Abate, CRA)
- RDA Wheat Data Interoperability Working Group (Esther Dzale Yeumo Kabore, INRA)
- Data Harmonization for Crop Simulation: the ICASA/AgMIP Approach (Jeffrey White, USDA)
- Value Chain Data (Christopher Brewster, Aston University)

Friday, 3 July 2015

Regions (Plenary)

- Agricultural Data Interoperability in Africa (Richard Kedemi)
- Semantics for Information Management in Agriculture: Embrapa's experience (Ivo Pierozzi Júnior, Embrapa, Brazil)
- Agricultural Data Interoperability in India (Devika Medalli, Institute of Statistics India)
- Mapping Agricultural Investments and Technologies (Melanie Bacou, IFPRI)

Breakout groups

1. Integrating semantic assets
2. Use cases
3. Tools

Discussion (Plenary)

Appendix C - Ideas from the pre-workshop questionnaire

Please describe your vision for semantics in agriculture and consider the following issues your response.

Please describe any problems your organization may be experiencing with regard to the interoperability of datasets or vocabularies. From your perspective, where are the bottlenecks, and what sorts of tools, resources, or actions are needed to solve them?

What do you see as the most pressing needs in agriculture for the coming decade? What sort of datasets are needed, and what sorts of vocabularies are needed to support access to and use of those datasets? Do particular areas need to be strengthened, such as integration of semantics in geographic information systems?

What priorities should the organizations represented at the workshop set for future actions? Are you aware of, or involved in, other relevant projects or initiatives in related areas? In what direction should we try to head over the coming decade?

Overall priorities

- KTBL: Need to Increase in resource efficiency, i.e., using **less inputs to produce higher outputs**.
- CNCP: Need to move from marginal utility to global sustainability, and integration of data is a means to achieve this. Need to improve our understanding of the **"components" of the agricultural landscape**.
- CIMMYT: Need to turn outputs of scientific data into **usable knowledge**.
- Alterra: To demonstrate added value, vocabularies must have stronger **links to real-world problems**.
- The way to achieve these goals is through **data, semantically enriched and connected. Curated, then exploited in applications.**

Role of semantics

- KALRO: Semantics plays a key role in **achieving interoperability**.
- Alterra: Semantics are the **glue for improved discovery** of data for use in applications.
- Syngenta: Accept that data production for agriculture will never be carried out in a single way. Semantics is the only way to **reconcile the diversity of agricultural systems to meet analytic needs**. A Semantic Web approach is the only way data producers can **break free from the shackles of vendor lock-in**.
- NAL: Need a **common vocabulary** for agriculture.
- KTBL: **Reuse** existing vocabularies.
- Aston: Semantic technologies also **integrate well with sensors** (for example in precision farming and agri-logistics) and the growing availability of **data from third parties** (for example concerning environmental impact). Semantic technologies can **reduce the effort and cost of making information available** across the supply chain.

Harmonization of vocabularies

- ARS: **Harmonization of vocabularies** is top priority, then tools to help people record and manage data. With harmonized vocabularies and tools, it is easier to move towards queryable and discoverable data resources.

Data usable by farmers and across disciplinary boundaries

- ISI: Data should be usable not just by IT professionals, but **scientists, field workers and researchers**.
- Alterra: Data should be usable to develop **new services** to farmers, processors, food chains, policy makers. Many applications stay in one discipline, so discovery **across disciplinary boundaries** would be huge step.

Existing vocabularies are not well known

- ASI: **Lack of awareness**: many who work in food systems do not even know that controlled vocabularies exist.
- INRA: Vocabularies are **hard to reuse** due to lack of visibility, poor description of goals, and lack of confidence by adopters. When vocabularies are hard to reuse, people **reinvent ad-hoc vocabularies**, which impedes interoperability.

The diversity of vocabularies and models makes them hard to map

- CIMMYT: **Mapping** to external vocabularies is difficult when **traits are measured by different methods and units**.
- Wageningen: For coming decade, we will have the challenge of **linking different models**, e.g., climate and crop growth.
- KTBL: **Coding systems vary** by agency and data producer - e.g., for crops. The coding systems are often **conceptually inconsistent** - e.g., mixing species with land use.
- INRA: Diversity of repositories means diversity of **legal issues**.
- ASI: When linking indicators to maps, lack of controlled vocabularies means **extra work and inconsistent interpretation**.
- IFPRI: Donors require **different reporting methods and vocabularies**, making it hard to map data for reuse.
- KTBL: We often develop tools and Linked Data server implementations to support **locally required features**.
- NAL: Landscape of information systems and databases is **fragmented**.
- INRA: We need methods for **linking new datasets to existing datasets**.
- Alterra: We need to link vocabulary, not through a super-vocabulary but by **aliasing the concepts and discovering gaps**.
- INRA: Need to see the big picture of **semantic coverage** in order to identify gaps.

Vocabularies can improve integration of the food supply chain

- Aston: Standardised vocabularies are essential for better **integrating food supply chains**, with their great diversity of actors (farmers, retailers, consumers...). A “need-to-know” attitude means information flows only “one-up, one down”. **Better tracking** can help respond to food emergencies.

Better search

- NAL: Data should be **searchable in multiple languages**.
- USDA: We need **natural-language** search engines that "understand" the meaning of searches; indexing vocabularies with richer **relationships**; **search engines** that take advantage of semantic relationships; and better **tools for semantic search and management**.
- We need to improve our **automated indexing tools** for entity and relationship extraction.
- Alterra: We need a **semantically enabled search engine**.
- NAL: Queries cannot be expressed in **natural language**.

Specific vocabularies needed

- USDA: We need vocabularies shared **across disparate information systems and datasets**.
- Bayer: We need vocabularies for integrating **taxonomy of species + varieties + European and Mediterranean Plant Protection Organization (EPPO)**; for typical **units** used in agricultural activities; for processes and activities related to **molecular breeding**.
- CIMMYT: Controlled vocabularies for **GIS data**.

- Alterra: We need vocabularies with variables commonly used in datasets addressing high-profile real-world problems, e.g., **monitoring and decision making**. Vocabularies have focused hitherto on metadata, not on **raw quantitative data**. Semantics should give users the means to **get data without understanding all the complexities**.
- Workshop: We need relationships that will **enrich the semantics** of the vocabulary.
- ASI: Controlled vocabularies for key terms in **certification standards** would be useful for consumers and large food companies.
- IFPRI: We should **standardize data collection and recording processes**, which are different in different countries.
- INRA: **Provenance** metadata is needed to contextualize and assess data. We need **guidelines for discovery metadata.**
- Cornell: A controlled vocabulary of **agriculture-related job titles**.

Tools, infrastructure, and formats

- INRA: We need **common formats for data exchange**; not just machine-readable identifiers, but human-friendly **naming conventions**; collaborative and intuitive **interfaces for describing resources** with published vocabularies.
- CIMMYT: We need **infrastructure to handle precision phenotyping images and genotype sequences.**
- We need services for **semantically annotating data** with ontologies and **systems for storing and exploiting ontology alignments and data annotations**.
- INRA: We need collaborative and intuitive **interfaces to find and integrate existing resources**.
- Cornell: We need accurate **mechanisms for disambiguating people** (authors of publications) and free-text organizational names (e.g., "FAO" vs "Food and Agricultural Organization"). With better geospatial data, relationships between people, areas of research, and events could be **viewed spatially**. We need better **tools for consuming data** -- Web services and APIs that can be easily used in a variety of platforms.

Data needed

- CNCP: We need better **data on soil** to improve sustainability and address climate change.
- CRA: For holistic agriculture, we need **soil maps**.
- ARS: Improving **data from field trials** is a key priority.
- KTBL: For less-input agriculture, need data on **variety trials, pests, nutrient supplies,** and **agricultural supplies.**
- CIMMYT: Need to link types of data: **agronomic data, socio-economic, climate, germplasm, geo-spatial, image, phenotypic, genotypic.**
- IFPRI: We need interoperable **information in agriculture, health, GIS, and soil** to see the big picture.
- CABI: We need **data at all geographical resolutions**.
- NAL: **Metadata about datasets** (descriptions) is poor.

Policy and culture, skills, training, incentives

- INRA: Need **data sharing policy**: what can be published, under what conditions.
- Bayer: follow the **"FAIR" principles (unique identifiers, persistent data, reusability…)**.
- KTBL: Integrating semantics in **GIS** systems is more about convincing people than a tech problem.
- CIMMYT: Need to change the **culture of knowledge sharing.**
- IFPRI: Lack of **staff training** and involvement in **curating data** and using normalized ontologies leads to individualized fixes.
- Workshop: We need **governance and quality control policies** to enable expansion of vocabulary 24/7 in

multiple languages.

- KALRO: Lack of **info management policies** makes it difficult to collect data for repositories. Stakeholders lack awareness of the value of semantics and open data. We need a platform to **sensitize stakeholders** on need to adopt tools and standards to make datasets accessible.
- INRA: Need **incentives** to induce scientists to provide provenance data
- NAL: Data producers **fear losing control** and misuse of "their" data.
- IFPRI: We need not just techies, but real **data curators**, including **GIS** skills. We need **staff training** in the use of tools, standards, and vocabularies.
- CABI: We need to address the importance of **manual curation** of vocabularies, which is labor-intensive and requires domain expertise.
- Alterra: We need **tutorials** on using vocabularies of standardized variables.
- ISI: Legacy data works fine within closed communities; interoperability becomes an issue for Web-wide use.
- KTBL: The hardest part is to **convince IT departments** re: globally distributed versus central, relational databases and to **convince management** that information is too valuable for survival of human race to keep locked up.
- Land Portal: There is a lack of understanding that opening data can make it more useful without de-possessing them of "ownership". We lobby for Linked Open Data as a way to unleash more potential than simply publishing "open data".

Use, reuse, upkeep

- ASI: We envision **continuously updating** our database with indicators and linking these indicators to controlled vocabularies.
- Bayer: **Vocabularies must evolve** with new technologies such as digital farming and precision phenotyping. The **long-term sustainability** of vocabularies is a problem.
- Workshop: We need to **distribute the workload among partners**.
- CRA: Re-use: **even data produced by a single organization** is not used in full complexity.
- CNCP: Build LOD applications in **collaboration among organizations**.
- IFPRI: We need to **balance broad coverage with enough precision** (e.g., Agricultural Technology Ontology).
- CABI: Whatever we do, long-term sustainability is a serious consideration. We need a viable **business model**.
- Cornell: Use GACS concepts to **link People with Organizations, Publications, and Events**.

Suggestions for concrete actions

- Bayer: We could **write a position paper** stating the needs of this community and encourage others to join.
- KTBL: A tool along the line of taginfo.osm.org (for Open Streetmap) that **crawls datasets** and provides stats on usage of RDF properties and classes.
- Alterra: A semantically-enabled data cube of **frequently used data sources**, available as public good. Would need to **incentivize users** to contribute their data to further extend the data cube.
- Syngenta: Join pockets of data collection in a **web reusable for food security**.
- ASI: **Collect success stories** of controlled vocabulary use in order to motivate groups to use them.
- CRA: **Smartphone apps** can promote data interoperability at level of organization (single point of access) and domain (GPS + camera for soil survey).
- KALRO: To address climate change, we need particular **geographic information systems** and data on **population density**.
- Alterra: We need **vocabulary-enabled applications** for real-world problems, from agricultural monitoring (e.g., existing GeoGLAM initiative) to information on climate and diseases.

Related projects

- CIMMYT: Planteome project.
- CIMMYT: Semantic GIS project.
- KALRO: WheatIS, International Wheat Information System.
- Bayer: Pistoia Alliance and other major successful initiatives (e.g. OpenPhacts, iPlant).
- COPO (Collaborative Open Plant Omics), to describe plant research datasets.
- Agroportal, which helps re-use biomedical research, e.g., with an ontology portal.
- Alterra: AgMIP ICASA variable list for crop models and agronomic trials.